Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 429629, 19 pages
http://dx.doi.org/10.1155/2014/429629

Research Article

Towards a Unified Sentiment Lexicon Based on Graphics Processing Units

Liliana Ibeth Barbosa-Santillán[1] and Inmaculada Álvarez-de-Mon y-Rego[2]

[1] *Department of Computer Science, Universidad de Guadalajara, Periferico Norte 799, Modulo L-308, Los Belenes, 45100 Guadalajara, JAL, Mexico*
[2] *Lingüistica Aplicada a la Ciencia y la Tecnología, Universidad Politécnica de Madrid, Madrid, Spain*

Correspondence should be addressed to Liliana Ibeth Barbosa-Santillán; lbarbosa@alumnos.fi.upm.es

Received 16 July 2013; Accepted 11 October 2013; Published 13 March 2014

Academic Editor: Yudong Zhang

This paper presents an approach to create what we have called a Unified Sentiment Lexicon (USL). This approach aims at aligning, unifying, and expanding the set of sentiment lexicons which are available on the web in order to increase their robustness of coverage. One problem related to the task of the automatic unification of different scores of sentiment lexicons is that there are multiple lexical entries for which the classification of positive, negative, or neutral $\{P, N, Z\}$ depends on the unit of measurement used in the annotation methodology of the source sentiment lexicon. Our USL approach computes the unified strength of polarity of each lexical entry based on the Pearson correlation coefficient which measures how correlated lexical entries are with a value between 1 and −1, where 1 indicates that the lexical entries are perfectly correlated, 0 indicates no correlation, and −1 means they are perfectly inversely correlated and so is the UnifiedMetrics procedure for CPU and GPU, respectively. Another problem is the high processing time required for computing all the lexical entries in the unification task. Thus, the USL approach computes a subset of lexical entries in each of the 1344 GPU cores and uses parallel processing in order to unify 155802 lexical entries. The results of the analysis conducted using the USL approach show that the USL has 95.430 lexical entries, out of which there are 35.201 considered to be positive, 22.029 negative, and 38.200 neutral. Finally, the runtime was 10 minutes for 95.430 lexical entries; this allows a reduction of the time computing for the UnifiedMetrics by 3 times.

1. Introduction

Written language has been the preferred medium of communication in order to express facts, assumptions, and opinions. The web has facilitated the connection of speakers overcoming the barriers imposed by location, language used, customs, context, culture, and so forth through electronic devices. Content producers are increasing their activity in blogs, web pages, portals, social networks, e-mails, chats, and so forth.

Surprisingly, the speaker boundaries of nationality are no longer distinguished, even using the Global Positioning System, due to the growth of a global migration and the merging of languages by multilingual professional communities.

In many countries, multilingual features occur in many families where parents have several mother tongues. They use a common language at home as they live in a country where another language is spoken. The children swap between an average of four languages simultaneously excluding the extra languages that they learn in school. For instance, according to the United States Census in 2007 [1], the percentage of individuals around five years and older who speak only English at home is 80.3% and who use a language other than English at home is 19.7%, among them Spanish accounts for 62.3% followed by Asian and Pacific Island languages with 15.0%.

On the other hand, the global economy is based on the digital applications such as e-commerce and online entertainment, social media including social enterprise, digital media advertising, the Internet, and cloud computing. For that reason, knowing the opinion of citizens becomes essential because they are increasingly active in the content production

which enterprises, researchers, government, and intelligence services consider that can be monitored.

To analyse the web to discover sentiment is a daunting task due to the difficulty of getting accurate results. Nevertheless, machine learning algorithms have obtained good results classifying sentiment within specialized domains and using controlled corpora.

Most existing studies have been conducted using cluster or statistic analysis in order to classify sentences, paragraphs, and documents. Furthermore, several rates and user profiles have been used in the collaborative assessment of services.

In summary, a collective interest is to understand the thought of a global society. In this context, structured linguistic resources are vital and they should be supported by a group of linguists working on a global level.

Lexicons are atomic linguistic resources necessary for processing information automatically. The web and the information explosion is making the available lexicons insufficient because the web is heterogeneous, multilingual, and dynamic. Even though there are approximations for automatically creating sentiment lexicons, they definitely should be improved in the areas of verification task and expert assessment.

Our research focuses on the four languages with the greatest number of first-language speakers including Chinese with 1,197 million, Spanish with 406 million, English with 335 million, and Portuguese with 202 million (based on Lewis, et al. [2]). Each family contains 15,000,000, 350,000, 88,000, and X words, respectively, according to the Cambridge, RAE, Oxford, and Grande Dicionário Houaiss da Língua Portuguesa dictionaries in 2013. Expecting to produce a 100% robust sentiment lexicon in our research is a titanic task. However, to automatically increase their robustness with improved quality is possible.

Another important challenge is the representation of lexically encoded knowledge and the researchers are suggesting new ways to do this. Moreover, their structures are different from each other, making it difficult to reuse, so their resources become a problem of interoperability.

In addition, the same lexical entry can be found in many sources having distinct assessments. In this case, the unification task is key since one of our goals is to compare the strength of polarity between sources of information and their symbols in several languages. However, the problem of unifying the strength of polarity is primarily a problem of processing power due to the size of the sentiment lexicon, which makes a hand-by-hand analysis simply not feasible.

If we want to know the correlation of a lexical entry to the rest of the Unified Sentiment Lexicon, then the number of possibilities is 9.08E + 009. So computing is a problem of time as the calculation involved is huge.

Lexical resources, especially those semantically annotated, are time consuming and require a lot of effort; thus, we tried to use as much already existing work as possible in our effort to build a Unified Sentiment Lexicon.

Sentiment Lexicons that our research used are

SL1 SentiWordNet developed by Istituto di Scienza e Tecnologie dell'Informazione,

SL2 Bing Liu Sentiment Lexicon developed by Illinois University,

SL3 MPQA lexicon developed by Pittsburgh University,

SL4 NTU Sentiment Dictionary (NTUSD) developed by the Institute of Information Science of Taiwan,

SL5 PanAmerican sentiment lexicon developed by the Polytechnic University of Madrid,

SL6 Spanish Travel Subjective Lexicon (STSL) developed by the Polytechnic University of Madrid.

Our research questions are

Q1 is it possible to unify the sentiment lexicons available on the web and align and expand them automatically?

Q2 Is it possible to transform a Unified Sentiment Lexicon into a generative lexicon based on a core ontology?

The following set of hypotheses covers the main features of the proposed solutions:

H1 the unification of sentiment lexicons allows for a robust linguistic resource,

H2 given different strengths of polarity of the same lexical entry, it is possible to compute a unified value,

H3 unification calculus of each one of the lexical entries with GPUs' local and global memory allow the reduction of hard disk access and increase processing speed.

Compared with previous work, the major contributions of this paper are as follows:

C1 a cluster of sentiment lexicons has been unified automatically and validated by experts,

C2 the Unified Sentiment Lexicon has been expanded with two more sentiment lexicons that were developed by our research group Communication in Specialized Domains,

C3 the task of unification uses parallel processing for computing each lexical entry with GPUs,

C4 USL compute was accelerated by 3 times in data processing,

C5 robustness of coverage of the Unified Sentiment Lexicon,

C6 a uniform representation of lexically encoded knowledge.

In summary, this paper describes the Unified Sentiment Lexicon (USL) approach for aligning, unifying, and expanding the sentiment lexicons available in an automatic way in order to increase their robustness of coverage obtaining as a result a large-scale Unified Sentiment Lexicon based on GPUs.

The remainder of the paper is structured as follows. Section 2 briefly presents the background of our work. Section 3 describes the USL approach. Section 4 describes how our USL approach was implemented and the different subtasks of the algorithm in detail. Section 5 presents details of our data sets, evaluation metrics, and the result. Finally, Section 6 presents our conclusions and future research.

2. Related Work

The related work considers the following: Section 2.1 data structures such as lexicons, specialized lexicons, and ontologies; Section 2.2 the methodologies available for building lexicons; and Section 2.3 the techniques of data processing focused on parallel processing and the kind of memory used for the TESLA architecture. Finally, we will present a summary table with the main features of the sentiment lexicons that are part of our study.

2.1. Data Structures. The lexical representation is founded in several data structures that form the basis of the linguistic resource, which is atomic. First, we will examine the lexicons; second, the generative lexicons; and finally, the ontologies.

2.1.1. Lexicons. According to Greame et al., a lexicon "is a list of words in a language along with some knowledge of how each word is used." A lexicon can have monolingual or multilingual properties. Lexicons are either created manually [3–5], semiautomatic [6, 7], or automatically [8–11]. When the lexicon is built manually a group of experts can annotate all the words in a specific corpus; the assessment is performed by consensus and each lexical entry is checked in order to achieve excellence.

On the other hand, automatic lexicons can be produced, based on a specific corpus, where the lexical entries included far exceed the total number that can be compiled manually. However, to assess their quality is not an easy task.

In state of the art research [3, 4, 8–10], each group examines a collection of documents and produces their own lexicon. As a result, we have a number of lexicons—some of them are available on the web and others are not—we describe those in Section 2.1.3. In fact, we believe that all work carried out by universities and research groups should be used in a homogeneous way. Furthermore, these lexicons available should be reused using algorithms that facilitate data processing.

The generative lexicon was introduced by Pustejovsky [12, 13] in 1995 with the aim of encoding selectional knowledge in language. Differently to a generative lexicon, an enumerative lexicon only includes the different senses for each lexical entry. In Pustejovsky's approach, there are four elements to encoding: lexical typing structure, argument structure, event structure, and qualia structure. For that reason Pustejovsky's model deals with (a) the knowledge representation of the lexicon and the relation between an object and its constituents, (b) the formal role that distinguishes the object within a larger domain, (c) the purpose and function of the object, and (d) the factors involved in the origin of an object; all these constitute the qualia structure.

Bergler [14] said that there are significant efforts involved in building and sharing a big generative lexicon that will be a standard in the scientific community.

2.1.2. Ontology. According to Gruber [15] "an ontology is an explicit specification of a conceptualization." In this sense, Graeme Hirst [16] said that "an ontology, as a nonlinguistic object that more directly represents the world, may provide an interpretation or grounding of word senses."

The following supported sentiment ontologies are available: Ontology-Supported Polarity Mining [17] introduced in 2005, it was based on an ontology for movie reviews, with the positive or negative polarity determined from a collection of texts and the Chinese Emotion Ontology based on HowNet [18] which contains just under 5,500 verb concepts covering 113 different emotion categories.

2.1.3. Sentiment Lexicons. Our research has focused on four sentiment lexicons that are available on the web: the National Taiwan University Sentiment Dictionary (NTUSD), SentiWordNet, Bing Liu Sentiment Lexicon, and the Subjectivity Lexicon of Theresa Wilson et al. (MPQA). These are explained below.

The National Taiwan University Sentiment Dictionary (NTUSD) [19] was developed by Lun-Wei Ku et al. It is based on the Chinese Network Sentiment Dictionary and the web. There are 11,088 terms that qualify for the simplified version, of which 2,812 are positive and 8,276 are negative. The reason for having the NTUSD was to identify positive and negative sentiment words and their weights in a corpus of blogs and news on the basis of Chinese word structures. Lun-Wei Ku et al. suggested a method for annotating 192 documents in order to tag them as positive, negative, or neutral on three levels: words, sentences, and documents. The results were that the F-measure was 73.18% and 63.75% for verbs and nouns, respectively. When the sentiment words were mined together with topical words, they achieved an F-measure of 62.16% at the sentence level and 74.37% at the document level.

SentiWordNet [20] is a lexical resource produced by Istituto di Scienza e Tecnologie dell'Informazione, Italy. The main objective is to automatically estimate the value of all entries of WordNet [21] as positive, negative, or neutral assigning to each a weight between zero and one according of the value. The reason to create the SentiWordNet automatically was that the WordNet has more than 155,287 entries and annotating them manually would be a time consuming task. The result is a sentiment lexicon with 117,659 terms classified into the same four lexical categories as WordNet: adjectives, nouns, verbs, and adverbs.

Bing Liu Sentiment Lexicon [22, 23] is a lexicon where a small list of adjectives was manually created and tagged with positive or negative labels. It is domain-independent and he proposed a technique to grow this list using WordNet. He used a web-mining method to obtain a set of adjectives in the same way that the speaker wrote them. Thus, their lexicon has entries that are not in the English dictionary. The results obtained are 2,006 positive words and 4,764 negative words.

The Subjectivity Lexicon of Theresa Wilson et al. [3] is a lexical resource where the words that are subjective in most

contexts are marked as being strongly subjective and those that may only have subjective usages in certain contexts are marked as weakly subjective. The process of building it was manual. The words were also classified according to their categories as nouns, verbs, adjectives, and adverbs. The results are 8,221 clues (as she call them) where 4,913 are negative, 2,721 are positive, and 571 are neutral.

The Spanish Travel Subjective Lexicon (STSL) [4] was built *ad hoc* based on a web-based corpus of blogs that were analysed within the framework of appraisal theory [24]. The blogs were analysed to create a subcorpus of sentences annotated according to appraisal and these sentences were classified as positive or negative considering some contextual rules that could influence the strength of the polarity. These sentences were used to build the lexicon. The words were classified according to their categories as nouns, verbs, adjectives, and adverbs; multiword units were also included. The result was 1,610 terms of which 857 are positive and 753 are negative.

The PanAmerican Sentiment Lexicon approach aims to classify according to polarity a set of internet resources focused on an event. The approach is based on four components: a crawler, a filter, a synthesizer, and a polarity analyzer. The main function of the crawler component is to search and find data from internet resources related to the event of interest. After locating the data, the filter component processes the data in order to remove noise. The filter component only debugs internet resources that are associated with the event. At this point, the corpus consists of large posts containing large amounts of data from many countries and in many languages. The synthesizer component represents the amount of data into clusters with similar expressions using unsupervised learning. Finally, the polarity analyzer component classifies each lexical entry as positive, neutral, or negative. The lexical categories are noun, adjective, verb, adverb. Finally, the result arrived at of 6083 positive, 5300 negative, and 5000 neutral.

2.2. Methodology. We have identified several kinds of methodology for building sentiment lexicons and have classified them as follows: automatic, semiautomatic, and manual.

(1) Automatic Methodology. First, the crawler is used to obtain a set of lexical terms in a controlled domain. Next, data preprocessing is performed and terms are assessed and classified as positive, negative, or neutral. The evaluation task involves using a subset of annotated lexical entries created manually by experts in order to measure the accuracy of the results. However, one of the limitations is the quality of the results because undertaking the evaluation task manually is not feasible. One of the advantages of this methodology is the higher number of terms produced compared to the results of the manual methodology.

(2) Semiautomatic Methodology. When linguistic resources use this methodology both manual and automatic annotations are used. An initial lexicon is annotated manually and this subset is used for training the algorithm, which will predict the level of matching for automatically classifying each new lexical category in the lexicon.

(3) Manual Methodology. Experts manually annotate each lexical entry in the appropriate lexical category. The lexicon quality is high, but the depth of the lexicon is less than that obtained with other methodologies.

2.3. Parallel Processing. Parallel processing [25] allows the running of several jobs at the same time and to accelerate the process by producing answers concurrently. Graphics Processing Units (GPUs) [26] allow the implementation of several algorithms [27].

These algorithms have been proved to provide acceleration from $2x$ to nx for specific problems [28].

Previous research has focused on image systems, simulations of fluids and molecular simulations with GPUS. Image systems [29] such as TechniScan uses ultrasonic waves to image the patient's chest in 20 minutes. Some other examples of the use of this technology are the following. The University of Cambridge was able to accelerate computational simulations of fluid dynamics [30] in order to perform rapid experimentation. Temple University performs molecular simulations [31] in order to reduce the environmental impact of detergents and cleaning agents. The time these simulation lasted was reduced from several weeks to a few hours.

We explored the use of GPU technology [32] in order to accelerate our data preprocessing.

TESLA [33] is a GPU architecture produced in 2003 by NVIDIA. It consists of a shared memory, constant cache, register file, double precision, Special Function Unit (SFU), Streaming Processor (SP), and a Warp Scheduler.

Memory is an essential component of high-performance computing. CUDA uses several types of memory [34] depending on the problem. The host memory is in the GPU system. CUDA provides APIs that enable faster access to the host memory by using the pager block and mapping the address direction on the GPU.

The memory device is in the GPU. It can be accessed by the dedicated memory controller. Data must be explicitly copied between the host memory and device memory. This memory can be organized and accessed in different ways [35].

Global memory can be static or dynamic. Access is by CUDA core pointers. Its main function is to translate the addresses.

Constant memory is read only. Its access is through a hierarchical cache optimized for transmission to several threads.

Local memory is stored in the stack: local variables which cannot be stored in records, parameters, and the return addresses of subroutines.

Texture memory is accessed by instructions for loading and storing. As well as constant memory, an independent cache is used in order to execute read only operations.

The GPU CUDA device is a multicore coprocessor. It is possible to log in through all the device memory without constraints. However, there will be variations in runtime according to the type of target memory.

Table 1: Sentiment lexicons that are available on the web and supported by universities.

Name	University	Positive	Negative	Neutral	Language	Methodology	Category	Order
Bing Liu	Illinois	2006	4783	0	English	Automatic	No	Alphabetic
MPQA	Pittsburg	2721	4913	571	English	Manual	N, V, Ad, Adv	Alphabetic
NTUSD	SINICA	2812	8276	0	Chinese	Semiautomatic	No	Phonetic
Pan American	UPM	6083	5300	5000	English Spanish Portuguese	Manual	N, V, Ad, Ad	Alphabetic
SentiWordNet	ISTI	857	753	0	English	Automatic	N, V, Ad, Ad	No
STSL	UPM	19619	29792	89135	Spanish	Manual	N, V, Ad, Ad Interjections diminutives phrases	Alphabetic

2.4. Summary Table. Table 1 summarizes the sentiment lexicons of our study according to several features such as name of the university where it was developed; depth of the lexical entries, which is the sum of all the positive, negative, or neutral entries; coverage of language; type of elaboration methodology; lexical categories; and ordering procedure.

3. The Approach Proposed

Our approach aims at aligning, unifying, and expanding the set of sentiment lexicons which are available on the web in order to increase their robustness of coverage. It is composed of ten components: FocusCrawlerEngine, SelectorLanguages, MetricSearcher, MetricTransformer, SentimentLexiconIntersection, LexicalEntries Substracter, LexicalEntriesDivisor, UnifiedMetrics, UnionSentimentLexiconEngine, and Lexicon2OntologyConverter.

The USL approach has as an input of sentiment lexicons which are supported by universities and which are available on the web. First, the SelectorLanguages component creates a group of sentiment lexicons according to their language and stores them in different knowledge bases. The MetricSearcher component performs an inspection of each one of the elements of the sentiment lexicons in order to identify if they have associated metrics. Then it saves the results in two knowledge bases: (a) MetricsLexicon and (b) NoMetricsLexicon. Next, the MetricTransformer component verifies if the metrics are not numerical in order to transform them with real values based on the original assessment. Consequently, our USL approach performs the intersection between all the sentiment lexicons. The common lexical entries are extracted with word, the strength of polarity, and sentiment lexicon values. Two knowledge bases are obtained as partial result: IntersectionLexicalEntries and NoIntersectionLexicalEntries.

The MetricsLexicon Knowledge base is the input for the LexicalEntries Substracter whose main function is to exclude all the IntersectionLexicalEntries. The USL approach is able to calculate a unified strength of polarity between all of the lexical entries of several sentiment lexicons. Thus, calculating the unified strength of polarity demands a high processing time because of the number of lexical entries. The LexicalEntriesDivisor split jobs in order to calculate the unification strength of polarity into balanced loads. The coprocessors compute the degree of unified subjective for each lexical entry. Its calculation is based on a previous assessment of the sentiment's lexical sources and the incomplete information. Each coprocessor produces a lexical knowledge base with the score of the unified metric. The UnionSentimentLexiconEngine unifies all the knowledge bases into one. As a result we have the Unified Sentiment Lexicon (USL). Finally, the Lexicon2OntologyConverter performs a transformation from data to Ontology Web Language (OWL).

The ten components of USL approach will be described in detail.

3.1. FocusCrawlerEngine. Their main function is to find the sentiment lexicons available on the web that would be supported by universities. Then we can define the following: (1) MPQA Lexicon, (2) Bing Liu Sentiment Lexicon, (3) SentiWordNet, (4) NTU Sentiment Dictionary, (5) Spanish Travel Subjective Lexicon, and (6) PanAmerican Games Sentiment Lexicon.

Consider

$$MPQALexicon : \doteq \{Type, Length, Word, Position, Stemed, PriorPolarity\}, \quad (1)$$

where

$Type = \{t \mid t$ is a string which measures the subjectivity degree of each clue$\}$,

$Length = \{l \mid l$ is a integer number that indicate the length of the clues$\}$,

$Word = \{w \mid w$ is a string with the token or stem of the clue$\}$,

$Position = \{p \mid p$ is a natural number than identify a lexical$\}$,

$Stemed = \{s \mid s$ is a boolean than should match all unstemmed variants$\}$,

$PriorPolarity = \{pp \mid pp$ is a string with the prior polarity of clue$\}$.

Consider

$$BingLiuSentimentLexicon : \doteq \{Words\}, \quad (2)$$

where Words $= \{w \mid w$ is a string with a word that is positive or negative$\}$.

Consider

$$Travel : \doteq \{NP, NN, AP, AN, VP, VN, adP, AdN\}, \quad (3)$$

where

$NP = \{np \mid np$ is a string with a word that is noun and positive$\}$,

$NN = \{nn \mid nn$ is a string with a word that is noun and negative$\}$,

$AP = \{ap \mid ap$ is a string with a word that is adjective and positive$\}$,

$AN = \{an \mid an$ is a string with a word that is adjective and positive$\}$,

$VP = \{vp \mid vp$ is a string with a word that is verb and positive$\}$,

$VN = \{vn \mid vn$ is a string with a word that is verb and positive$\}$,

$AdP = \{adp \mid adp$ is a string with a word that is adverb and positive$\}$,

$AdN = \{and \mid adn$ is a string with a word that is adverb and positive$\}$.

Consider

$$NTUSD : \doteq \{Words\}, \quad (4)$$

where Words $= \{w \mid w$ is a string with a word that is positive or negative$\}$.

Consider

$$SentiWordNet : \doteq \{POS, ID, PosScore, NegScore, SynsetTerms\}, \quad (5)$$

where

$POS = \{p \mid p$ is a character of WordNet$\}$,

$ID = \{id \mid id$ is a character of WordNet$\}$,

$PosScore = \{ps \mid ps$ is a real number with the positive score assigned by SentiWordNet$\}$,

$NegScore = \{ns \mid id\ ns$ a real number with the negative score assigned by SentiWordNet$\}$,

$SynsetTerms = \{st \mid st$ is a string number with the terms$\}$.

Consider

$$PanAmericanSL : \doteq \begin{cases} ID, Timestamp, Word, Postive, \\ Negative, Neutral, \\ Noun, Adjective, Verb, \\ Adverb, Language, \end{cases} \quad (6)$$

where

$ID = \{id \mid id$ is an integer number to identify a word$\}$,

$TimeStamp = \{ts \mid ts$ is a date of assessment$\}$,

$Word = \{w \mid w$ is a string with the word$\}$,

$Positive = \{p \mid p$ is a boolean with 1 if it is positive$\}$,

$Negative = \{n \mid n$ is a boolean with 1 if it is negative$\}$,

$Neutral = \{nt \mid nt$ is a boolean with 1 if it is neutral$\}$,

$Noun = \{noun \mid noun$ is a boolean with 1 if it is noun$\}$,

$Adjective = \{ad \mid ad$ is a boolean with 1 if it is adjective$\}$,

$Verb = \{v \mid v$ is a boolean with 1 if it is verb$\}$,

$Adverb = \{ad \mid ad$ is a boolean with 1 if it is adverb$\}$,

$Language = \{l \mid l$ is a string with the name of the language$\}$.

3.2. SelectorLanguages. This identifies the language in a subset of lexical entries in order to search for specific words in four languages: Chinese, Spanish, English, and Portuguese. The result is the cluster of sentiment lexicons arranged by language, as shown in (1).

Consider

$$LanguageWords \subseteq LexicalEntries \longleftrightarrow (\forall x)\left(x \in LanguageWords \longmapsto x \in LexicalEntries\right). \quad (7)$$

3.3. MetricSearcher. It is responsible for selecting the strength of polarity label of each of the sentiment lexicons clusters. For example, "PriorPolarity," "PosScore," "NegScore," and "ScoreSubjectivity," among others. Besides, it searches for strength of polarity and indicates whether the values are numerical or nominal. It splits its result in two knowledge bases: MetricsLexicon and NoMetricsLexicon.

Consider

$$\begin{aligned} MetricsLexicon &= \{x \mid (x \in LexicalEntries) \\ &\qquad \wedge (x \notin NoMetrics)\}, \\ NoMetricsLexicon &= \{y \mid (y \in LexicalEntries) \\ &\qquad \wedge (y \notin Metrics)\}. \end{aligned} \quad (8)$$

3.4. MetricTransformer. The MetricTransformer works by transforming the strength of polarity nominal value into the real value of each sentiment lexicon. It has two variables: type and pos. Type can take two values: $strongsub = .9$ and $weaksubj = .5$. Pos can take four values: $adj = .9$, $verb = 1$, $adverb = .8$, and $noun = .7$. The new strength of polarity is the multiplication between the two variables.

Consider

$$NewStrengthPolarity = type * pos. \quad (9)$$

3.5. SentimentLexiconIntersection. This component compiles with the intersection for all the lexical entries of each sentiment lexicon cluster. It aims to identify which lexical entries appear more than once in order to select them for processing.

Therefore, the two knowledge bases are IntersectionLexicalEntries and NoIntersectionalLexicalEntries. For example, the cluster of sentiment lexicons grouped by English language has four elements *EnglishCluster* = {*PanAmericanSentimentLexicon, BingLiuSentimentLexicon, SentiWordNet, and MPQALexicon*}. These intersections are shown in (10).

Consider

$$\begin{aligned}
&SentiWordNet \cap BingLiu = \{x \mid (x \in SentiWordNet) \\
&\qquad \wedge (x \in BingLiu)\}, \\
&SentiWordNet \cap MPQA = \{x \mid (x \in SentiWordNet) \\
&\qquad \wedge (x \in MPQA)\}, \\
&MPQA \cap BingLiu \\
&\quad = \{x \mid (x \in MPQA) \wedge (x \in BingLiu)\}, \\
&SentiWordNet \cap BingLiu \cap MPQA \\
&\quad = \{x \mid (x \in SentiWordNet) \wedge (x \in BingLiu) \\
&\qquad \wedge (x \in MPQA)\}, \\
&PanAmerican \cap SentiWordNet \\
&\quad = \{x \mid (x \in PanAmerican) \wedge (x \in SentiWordNet)\}, \\
&BingLiu \cap PanAmerican \\
&\quad = \{x \mid (x \in BingLiu) \wedge (x \in PanAmerican)\}, \\
&PanAmerican \cap SentiWordNet \cap BingLiu \\
&\quad = \{x \mid (x \in PanAmerican) \\
&\qquad \wedge (x \in SentiWordNet) \wedge (x \in BingLiu)\}, \\
&PanAmerican \cap MPQA \\
&\quad = \{x \mid (x \in PanAmerican) \wedge (x \in MPQA)\}, \\
&SentiWordNet \cap PanAmerican \cap MPQA \\
&\quad = \{x \mid (x \in SentiWordNet) \wedge (x \in PanAmerican) \\
&\qquad \wedge (x \in MPQA)\}, \\
&PanAmerican \cap BingLiu \cap SentiWordNet \cap MPQA \\
&\quad = \{x \mid (x \in PanAmerican) \wedge (x \in BingLiu) \\
&\qquad \wedge (x \in SentiWordNet) \wedge (x \in MPQA)\}.
\end{aligned} \tag{10}$$

3.6. LexicalEntriesSubstracter. This gets the rest of all the elements that have been assessed by each university. It subtracts MetricsLexicon from IntersectionLexicalEntries.

Consider

$$\begin{aligned}
&NoIntersectionLexicon \\
&\quad = InitialLexicon(x) - IntersectionLexicon, \\
&NoIntersectionalLexicon \\
&\quad = def\ \{s \mid s \in InitialLexicon, \\
&\qquad s \notin IntersectionLexicon\}.
\end{aligned} \tag{11}$$

3.7. LexicalEntriesDivisor. It has as its input the intersection of all the lexical entries. It divides the knowledge base into equal parts for processing.

Consider

$$N = \frac{LexicalEntriesTotal}{NumberCores}. \tag{12}$$

3.8. UnifiedMetrics. This performs an estimate of each lexical entry of the IntersectionLexicon in order to predict its value. There are two procedures: (1) $UnifiedMetrics_{CPU}$ and (2) $UnifiedMetrics_{GPU}$.

$UnifiedMetrics_{CPU}$ uses a Pearson correlation formula as shown in (13) applied between the Unified Sentiment Lexicon and each of the sentiment Lexicons by cluster.

$UnifiedMetrics_{GPU}$ algorithm in detail is explained in Section 4.

Consider

$$r = \frac{\sum XY - \sum X \sum Y/N}{\sqrt{\left(\sum x^2 - \left(\sum x^2\right)/N\right)\left(\sum y^2 - \left(\sum y^2\right)/N\right)}}. \tag{13}$$

3.9. UnionSentimentLexiconEngine. Its function is to join all the result knowledge bases of the coprocessors together and as output the Unified Sentiment Lexicon is obtained.

Consider

$$(\forall x)(\exists y)(\forall z)\left[z \in y \longleftrightarrow (\exists t)(t \in x \wedge z \in t)\right], \tag{14}$$

where $UnifiedSentimentLexicon_1 \bigcup \ldots \bigcup \ldots$. $UnifiedSentimentLexicon_n \leftarrow x \in this \Leftrightarrow x \in UnifiedSentimentLexicon_1$ *or* $x \in UnifiedSentimentLexicon_n$.

3.10. Lexicon2OntologyConverter. Their main function is to transform the Unified Sentiment Lexicon into a Domain Ontology: OntoLexicon as defined as follows.

3.10.1. OntoLexicon. The OntoLexicon Ontology is a conceptual description based on a lexicon of the subjective words in Natural Language as shown in (15). The OntoLexicon Ontology consists of four disjoint sets C, R, A, and τ, where C means concept identifiers (16), R means relation identifiers (17) and (18), A means attribute identifiers (19), and τ means data types (20).

Consider

$$OntoLexicon := (C, \leq c, R, \gamma_R, \leq_R, A, \gamma_A, \tau). \tag{15}$$

The set C of concepts is

$$C :\doteq \begin{cases} Adjectives, NegativeAdjectives, \\ PositiveAdjectives, Adverbs, NegativeAdverbs, \\ PositiveAdverbs, Verbs, NegativeVerbs, \\ PositiveVerbs, Nouns, NegativeNouns, \\ PositiveNouns. \end{cases} \tag{16}$$

The set R of relations is

$$R :\doteq \begin{cases} entry\_of, document\_of, paragraph\_of, \\ sentence\_of, adverb\_in, articles\_in, \\ prepositions\_in, nouns\_in, adjectives\_in, \\ verbs\_in, subject\_of, \\ predicate\_of, \end{cases} \tag{17}$$

where the relation hierarchy defines that Lexical has the relation entry_of that belongs to SentimentLexicon. Corpora has the relation document of that belongs to documents, following the same logic where the rest of the relations are defined as

$$\begin{aligned}
\gamma R\left(entry_of\right) &= \left(Lexical, SentimentLexicon\right),\\
\gamma R\left(document_of\right) &= \left(Documents, Corpora\right),\\
\gamma R\left(paragraph_of\right) &= \left(Paragraphs, Documents\right),\\
\gamma R\left(sentence_of\right) &= \left(Sentences, Paragraphs\right),\\
\gamma R\left(adverbs_in\right) &= \left(Adverbs, Sentences\right),\\
\gamma R\left(articles_in\right) &= \left(Articles, Sentences\right),\\
\gamma R\left(prepositions_in\right) &= \left(Prepositions, Sentences\right),\\
\gamma R\left(nouns_in\right) &= \left(Nouns, Sentences\right),\\
\gamma R\left(adjectives_in\right) &= \left(Adjectives, Sentences\right),\\
\gamma R\left(verbs_in\right) &= \left(Verbs, Sentences\right),\\
\gamma R\left(subject_in\right) &= \left(Subjects, Sentences\right),\\
\gamma R\left(predicate_in\right) &= \left(Predicates, Sentences\right).
\end{aligned} \tag{18}$$

The set A of attribute identifiers is

$$A :\doteq \begin{cases}
sentimentlexicon, author, strengthofpolarity,\\
paragraph, sentence, subject,\\
predicate, article, noun, nounN,\\
nounP, verb, verbN, verbP,\\
adverb, adverbN, adverbP, adjective,\\
adjectiveN, adjectiveP.
\end{cases} \tag{19}$$

The set τ of datatypes contains only one element, a string, is shown

$$\tau := \left(string\right). \tag{20}$$

The first axiom defines the concept NegativeAdverbs as equivalent to saying that there is a negative adverb which stands in a *adverb_in* relation with the corresponding sentence, following the same logic where the rest of the axioms are defined as

$$\begin{aligned}
&\forall x \left(NegativeNouns\left(x\right) \longleftrightarrow \exists y \wedge noun_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(PositiveNouns\left(x\right) \longleftrightarrow \exists y \wedge noun_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(NegativeAdjective\left(x\right) \longleftrightarrow \exists y \wedge adjective_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(PositiveAdjective\left(x\right) \longleftrightarrow \exists y \wedge adjective_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(NegativeAdverbs\left(x\right) \longleftrightarrow \exists y \wedge adverb_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(PositiveAdverbs\left(x\right) \longleftrightarrow \exists y \wedge adverb_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(NegativeVerbs\left(x\right) \longleftrightarrow \exists y \wedge verb_in\left(x, y\right) \wedge Sentences\left(y\right)\right),\\
&\forall x \left(PositiveVerbs\left(x\right) \longleftrightarrow \exists y \wedge verb_in\left(x, y\right) \wedge Sentences\left(y\right)\right).
\end{aligned} \tag{21}$$

4. Algorithm in Detail

The input of a Unified Sentiment Lexicon (USL) approach consists of the sentiment lexicons that are available on the web and supported by universities. The USL approach then processes all of them. The result is the Unified Sentiment Lexicon (USL). Here, we will describe the algorithm in detail.

The first step is to group the sentiment lexicons into clusters by language as following:

$$\begin{aligned}
x &= \{ChineseLanguage, SpanishLanguage,\\
&\qquad EnglishLanguage, PortugueseLanguage\},\\
Cluster_{x1} &= \{SentimentLexicon_1, \ldots,\\
&\qquad SentimentLexicon_n\}.
\end{aligned} \tag{22}$$

The second step is to search lexical entries that have been assessed by each sentiment lexicon. In the assessment task, some authors and their methods have used nominal values, while others have used real values. If they are linguistic values, then the USL approach transforms them into real values. There must then be an intersection of lexical entries in at least two sentiment lexicons in order to unify the strength of polarity of several sentiment lexicons into one.

Following this, our approach calculates the Pearson correlation score between each sentiment lexicon and the USL by obtaining as many constants as there are sentiment lexicons in the cluster. For example, if the cluster belongs to the English language, then there are four constants that fall into each sentiment lexicon, as shown in the Pearson correlation set $PearsonCorrelation = \{p_1, p_2, p_3, p_4\}$. This calculation is performed only once and executed by the CPU.

Since the number of lexical entries is high, the computation of the USL score should be divided into several coprocessors (cores) in order to accelerate the process. In fact, each coprocessor of the GPU has as an input: (a) the strength of polarity of n lexical entries and (b) the vector with Pearson values. Each coprocessor computes the strength of polarity of every lexical entry until there are no lexical entries left. The score for each lexical entry is multiplied by the Pearson correlation between all the sentiment lexicons, as shown in (23) and Table 2.

Consider

$$\begin{aligned}
\alpha_i &= p_1 * v_i,\\
\beta_i &= p_2 * w_i,\\
\gamma_i &= p_3 * y_i,\\
\delta_i &= p_4 * z_i.
\end{aligned} \tag{23}$$

In addition, USL performs a total of subjectivity sums, as shown in (24) and Table 2.

Consider

$$\varepsilon_i = \alpha_i + \beta_i + \gamma_i + \delta_i. \tag{24}$$

Table 2: The process to calculate the USL strength of polarity.

Words	Sentiment Lexicon$_1$	Sentiment Lexicon$_2$	Sentiment Lexicon$_3$	Sentiment Lexicon$_n$	α	β	γ	δ	ϵ	ζ	USL
Word$_1$	v_1	x_1	y_1	X	α_1	β_1	γ_1	δ_1	ϵ_1	ζ_1	$\mathbf{usl}_1 = \epsilon_1/\zeta_1$
Word$_2$	X	x_2	y_2	z_2	α_2	β_2	γ_2	δ_2	ϵ_2	ζ_2	$\mathbf{usl}_2 = \epsilon_2/\zeta_2$
Word$_3$	X	x_3	y_3	z_3	α_3	β_3	γ_3	δ_3	ϵ_3	ζ_3	$\mathbf{usl}_3 = \epsilon_3/\zeta_3$
Word$_4$	v_4	x_4	y_4	z_4	α_4	β_4	γ_4	δ_4	ϵ_4	ζ_4	$\mathbf{usl}_4 = \epsilon_4/\zeta_4$
Word$_{\dots}$	$v_{\dots}$	$x_{\dots}$	$X_{\dots}$	$z_{\dots}$	$\alpha_{\dots}$	$\beta_{\dots}$	$\gamma_{\dots}$	$\delta_{\dots}$	$\epsilon_{\dots}$	$\zeta_{\dots}$	$\mathbf{usl}_{\dots} = \epsilon_{\dots}/\zeta_{\dots}$
Word$_n$	v_n	X	y_n	z_n	α_n	β_n	γ_n	δ_n	ϵ_n	ζ_n	$\mathbf{usl}_n = \epsilon_n/\zeta_n$

```
(1) procedure UnifiedSentimentLexicon(seeds)
(2)    SentimentLexicons ← FocusCrawlerEngine(seeds);
(3)    Clusters ← SelectorLanguages(SentimentLexicons);
(4)    for i ← 1, NumberofClusters do
(5)       for j ← 1, NumberofSentimentLexicons do
(6)          for k ← 1, NumberofLexicalEntries do
(7)             if MetricSearcher(LexicalEntry(k)) = 1 then
(8)                MetricsLexicon(j)(k) ← LexicalEntry(k);
(9)             else
(10)               NoMetricsLexicon(j)(k) ← LexicalEntry(k);
(11)            end if
(12)            if MetricTransformer(NoMetricsLexicon(j)(k))≥0 then
(13)               MetricsLexicon(j)(k) ← NoMetricsLexicon(j)(k);
(14)            end if
(15)         end for
(16)      end for
(17)   end for
(18)   if SentimentLexiconIntersection(Clusters(MetricsLexicon)) = 1 then
(19)      IntersectionLexicalEntries(j)(k) ← Clusters(MetricsLexicon);
(20)   else
(21)      NoIntersectionLexicalEntries(j)(k) ← Clusters(MetricsLexicon);
(22)   end if
(23)   NoIntersectionLexicalEntries(j)(k) ← LexicalEntriesSubstracter
(24)   (SentimentLexicon, IntersectionLexicalEntries);
(25)   n ← LexicalEntriesDivisor(IntersectionLexicalEntries, NumberofCoresGPU);
(26)   for i ← 1, NumberofClusters do
(27)      for j ← 1, n do
(28)         r ← UnifiedMetrics_CPU(Cluster(i)(SentimentLexicons));
(29)         USL(i) ← UnifiedMetrics_GPU(n, r, Cluster(i)(InstersectionLexicalEntries));
(30)         UnifiedSentimentLexicon ← UnionSentimentLexiconEngine(j);
(31)      end for
(32)   end for
(33)   OntoLexicon ← Lexicon2OntologyConverter(UnifiedSentimentLexicon);
(34) end procedure
```

Algorithm 1: The main USL approach.

The USL score is normalized by dividing the total number of subjectivity for each lexical entry by the Pearson correlation sum of the lexical entries that were assessed $\zeta_1 = p_1 + p_2 + p_3 + p_4$, as follows:

$$USL_1 = \frac{\varepsilon_1}{\zeta_1}. \tag{25}$$

The GPU results are the lexical entries combined with the USL score (these are input by the CPU). Their main function is to join all the partial results in the USL.

Finally, the CPU transforms the USL into an ontology called OntoLexicon in OWL language.

The pseudocode of the main USL approach is shown in Algorithm 1, and some of the procedures of the USL approach are shown in Algorithm 2.

5. Experimental Details and Performance Results

The following section includes a detailed description of how the experiment was conducted. The first part describes the

```
(1) procedure FocusCrawlerEngine(Seeds)
(2)    for i ← 1, SizeSeeds do
(3)       SentimentLexicons(i) ← Download(Lexicon(i));
(4)    end for
(5) end prcedure
(6) procedure SelectorLanguages(SentimentLexicons)
(7)    for i ← 1, NumberofSentimentLexicons do
(8)       switch Language(SentimentLexicon(i)) do
(9)          case English
(10)            assert(Cluster1 ← SentimentLexicon(i))
(11)         case Spanish
(12)            assert(Cluster2 ← SentimentLexicon(i));
(13)         case Chinese
(14)            assert(Cluster3 ← SentimentLexicon (i));
(15)         case Portuguese
(16)            assert(Cluster4 ← SentimentLexicon(i));
(17)   end for
(18) end procedure
(19) procedure MetricSearcher(LexicalEntry)
(20)    if LexicalEntry ≥ 0 then
(21)       x ← 1;
(22)    else
(23)       x ← 0;
(24)    end if
(25) end procedure
(26) procedure MetricTransformer(NoMetricsLexicon)
(27)    NoMetricsLexicon(newMetric) ← NoMetricsLexicon(type)∗
     NoMetricsLexicon(pos);
(28) end procedure
(29) procedure SentimentLexiconIntersection(Clusters(MetricLexicon))
(30)    for i ← 1, NumberofCluster do
(31)       for j ← 1, NumberofLexicalEntries do
(32)          if MetricsLexicon(Cluster(i)(LexicalEntry(j)))
(33) = MetricsLexicon(Cluster(i + 1)(LexicalEntry(j))) then
(34)             intersection ← 1;
(35)          else
(36)             intersection ← 0;
(37)          end if
(38)       end for
(39)    end for
(40) end procedure
(41) prcedure LexicalEntriesSubstracter(SentimentLexicon, IntersectionLexicon)
(42)    LexiconSubstracter ← SentimentLexicon − IntersectionLexicon;
(43) end procedure
(44) prcedure LexicalEntriesDivisor(LexicalEntriesTotal, NumberofCores)
(45)    for i ← 1, NumberofCluster do
(46)       n ← LexicalEntriesTotal/NumberofCores;
(47)    end for
(48) end procedure
```

Algorithm 2: Some of the procedures of USL approach.

objectives of the experiment, and the second part focuses on the results obtained after the experiment was conducted.

The experimental setup had two objectives: to unify the sentiment lexicons and to carry out an analysis of our results by expert linguists.

The analysis was carried out in two ways: (1) where the USL approach ran automatically and (2) a linguistic evaluation of the quality of a subpart of the results obtained. The first task was to obtain sentiment lexicons available on the web validated by universities using the FocusedCrawlerEngine and two sentiment lexicons developed by our research group Communication in Specialized Domains were added.

The knowledge base of sentiment lexicons has been described in Section 3.1.

Figure 1 shows a partial view of the content of each of them: the Bing Liu sentiment lexicon appears in two files,

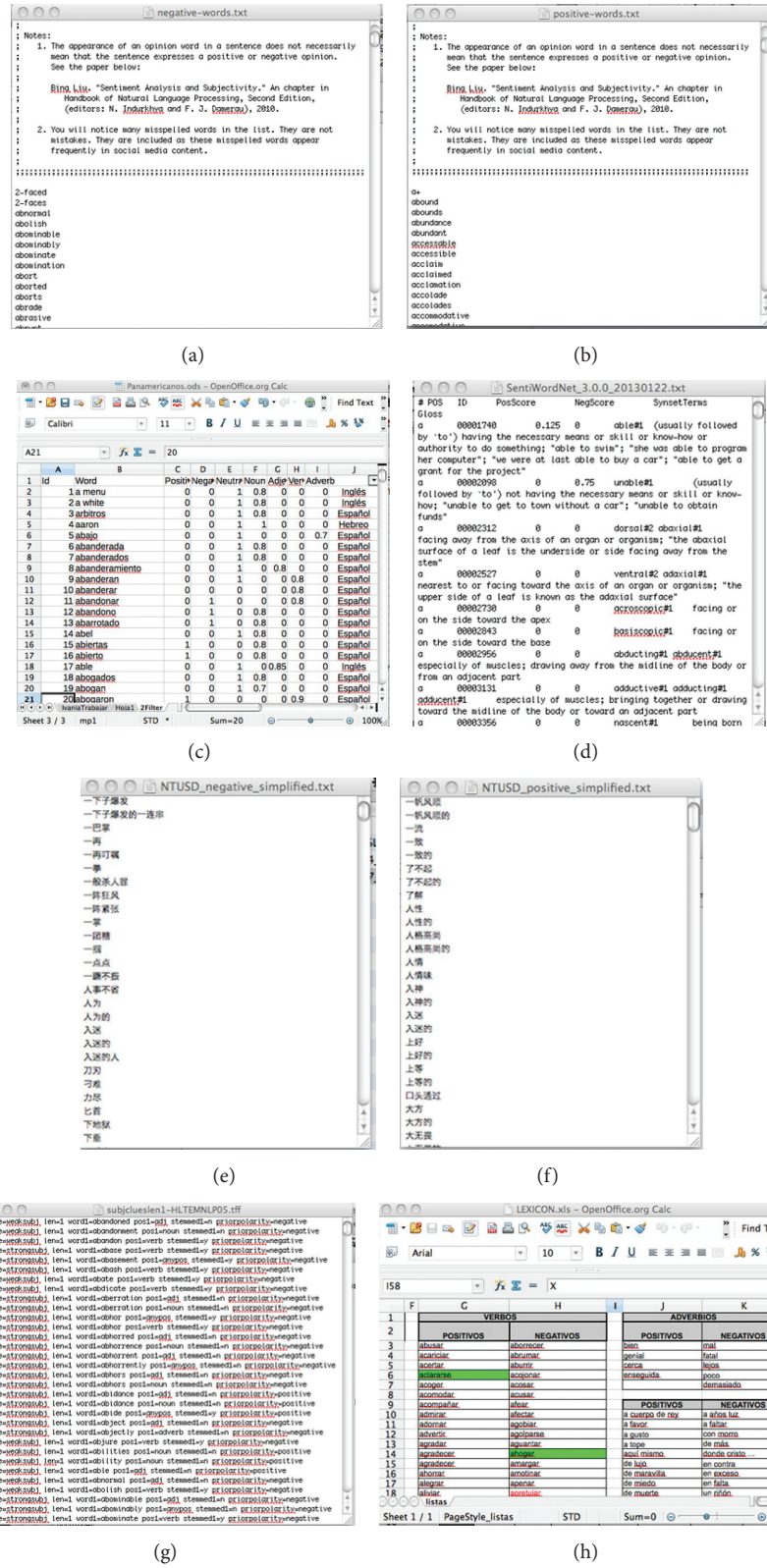

Figure 1: Input of USL approach.

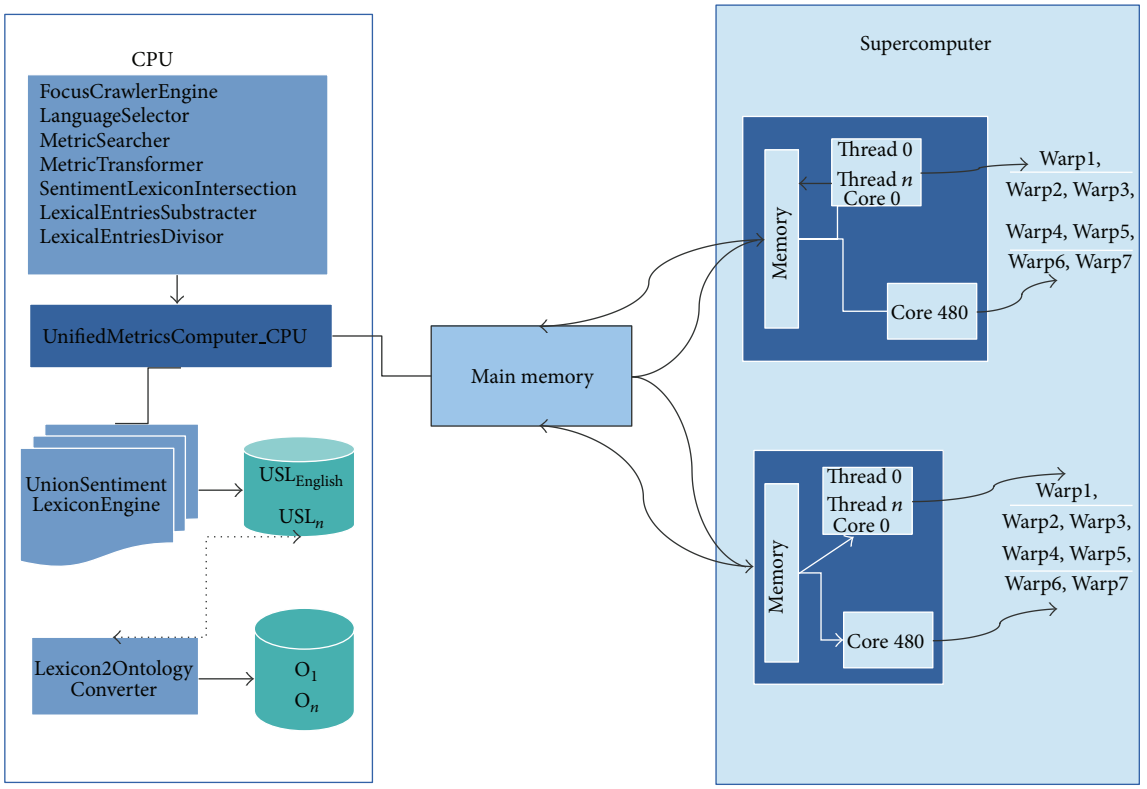

Figure 2: Architecture of USL approach.

as shown in Figure 1(a) and Figure 1(b); the Pan American sentiment lexicon is in Figure 1(c); SentiWordNet is in Figure 1(d); NTU Sentiment Dictionary separates into positive and negative, as shown in Figures 1(e) and 1(f); MPQA lexicon is in Figure 1(g); and Spanish Travel Subjective Lexicon is in Figure 1(h).

The second task was to filter those lexical entries that do not appear at least in two of the sentiment lexicons. The process has been described in Section 4, and an architectural view of the USL is displayed in Figure 2.

As a result, we obtained a total number of lexical entries for each sentiment lexicon: Bing Liu sentiment lexicon has 6789; MPQA lexicon contains 8221; NTU Sentiment Dictionary has 11088; Pan American has 16383; SentiWordNet has 111,711; and Spanish Travel Subjective Lexicon has 1610. The rate of each of them is displayed in Figure 3.

SentiWordNet is the lexicon with the highest number of lexical entries marked with positive polarity (12080 lexical entries); Spanish Travel Subjective Lexicon has the lowest (857 lexical entries), as shown in Figure 4(a). Bing Liu sentiment lexicon has 2006; MPQA lexicon has 2721; NTU Sentiment Dictionary has 2812; and Pan American sentiment lexicon has 6083.

In the case of neutral polarity, SentiWordNet stands out in the figure because an important subset of lexical entries (88564) has 0.0 as a strength of polarity. However, not all the sentiment lexicons assessed have neutral polarity, for example, Bing Liu sentiment lexicon, Spanish Travel Subjective Lexicon, and NTU Sentiment Dictionary; MPQA has 571

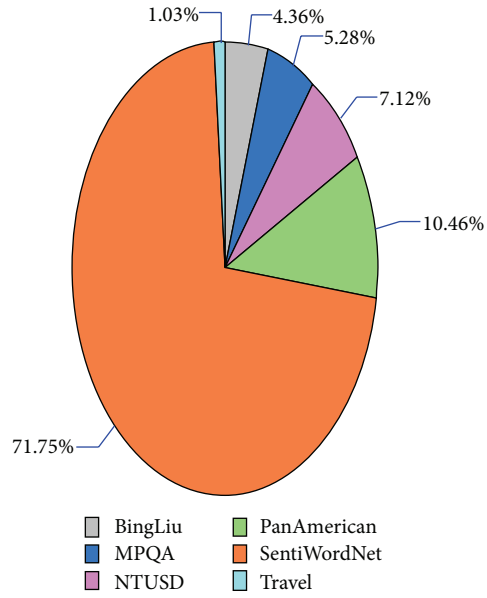

Figure 3: Rate of lexical entries total by sentiment lexicon.

and PanAmerican sentiment lexicon has 5000, as shown in Figure 4(b).

Figure 4(c) shows that, for negative polarity, SentiWordNet again has the highest number with 11067 lexical entries and again Spanish Travel Subjective Lexicon has the lowest

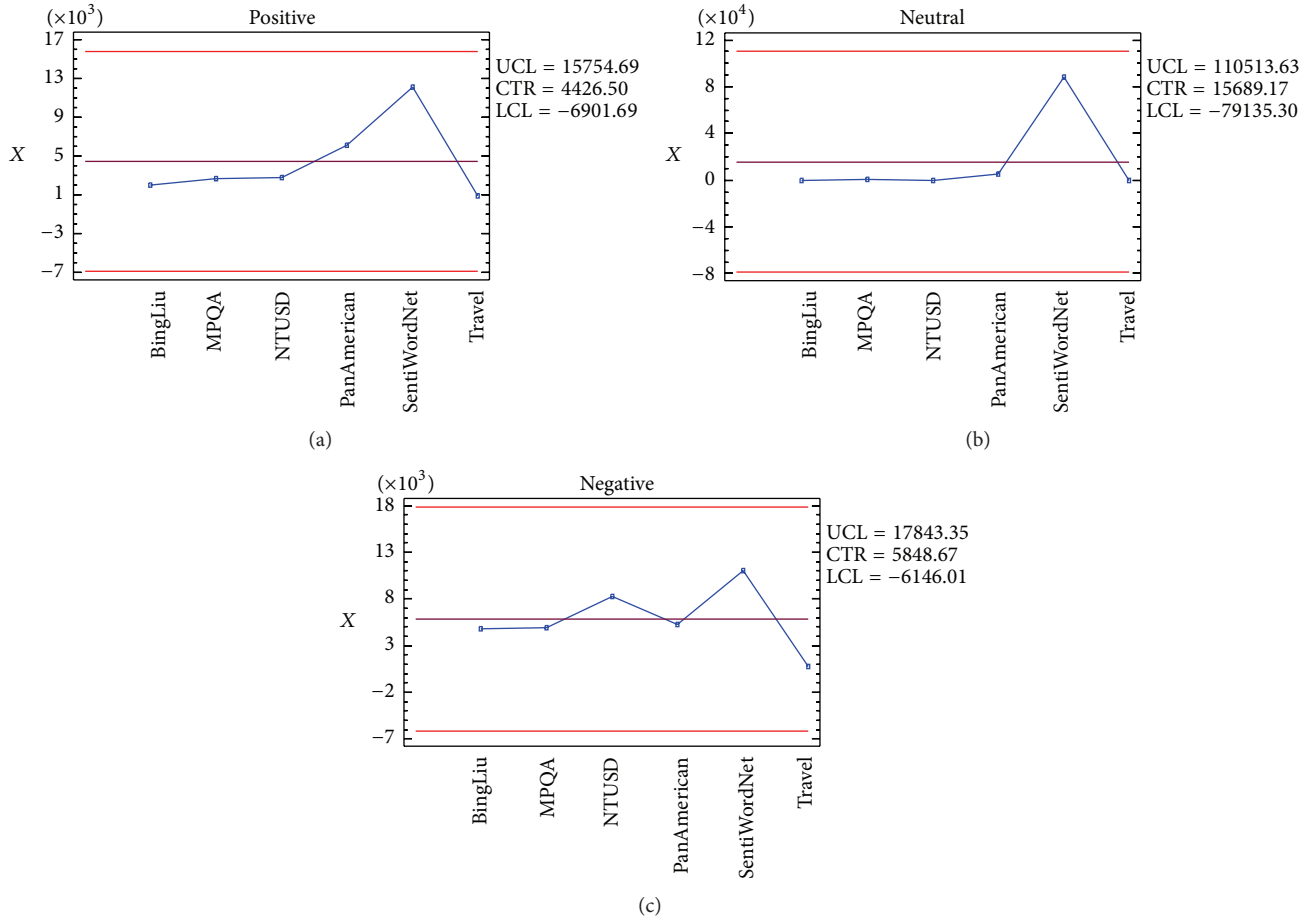

Figure 4: Number of lexical entries by polarity category for all the sentiment lexicons.

with 753. Bing Liu sentiment lexicon has 4783; MPQA lexicon has 4913; NTU Sentiment Dictionary has 8276; and Pan-American has 5300.

A partial result is shown in Tables 3, 4, and 5, where each table corresponds to a cluster for each language.

For the first cluster—English—a subset of lexical entries is shown in Table 3 and Figure 5.

Table 3 presents the lexical entry in each of the sentiment lexicons, the processing, and in the final column the strength of polarity of the USL.

The results are quite satisfactory although some minor problems have been detected. These problems are mainly due to the existence of expressions that can have both a positive and a negative value, and only one of the values is signalled. In the subset analysed, for instance, that is the case of the word *basic*, which can sometimes have a negative value when it is used to refer to the attributes or properties of an object or to the quality or level as in "the hotel room was too *basic*." Another problem is the influence of the results of considering all the lexicons for the final result. That is what happens in the case of the word achievement. The word is correctly classified as positive, but the degree of positiveness is too low due to the fact that in SentiWordNet evaluates it with zero, thus diminishing the final score.

Table 4 shows two sentiment lexicons, STSL and Pan-American. These two lexicons have lexical entries in common, however, STSL has not assessed the strength of polarity since it has classified them simply as positive or negative. For that reason, after processing the data, the USL score obtains the strength of polarity corresponding to the only sentiment lexicon that has been assessed.

For the Chinese cluster in the first attempt of alignment with the English language a problem arose because there is not direct correspondence as different English words are represented by the same Chinese symbols.

In addition only one Chinese sentiment lexicon has been found and USL approach tries to align it with SentiWordNet but this is only for purposes of exemplification because the meanings and terms in each language are different as shown in Table 5. For example, the English word *courageous* and the Chinese word do not have exactly the same meaning. Its Chinese translation is an idiom which means "fuly satisfaction." Another case is the word *severe* which in Chinese can only be express by using the word meaning "*strict*."

Another result of USL approach is OntoLexicon that was implemented in OWL language, a portion of which is shown in Algorithm 3.

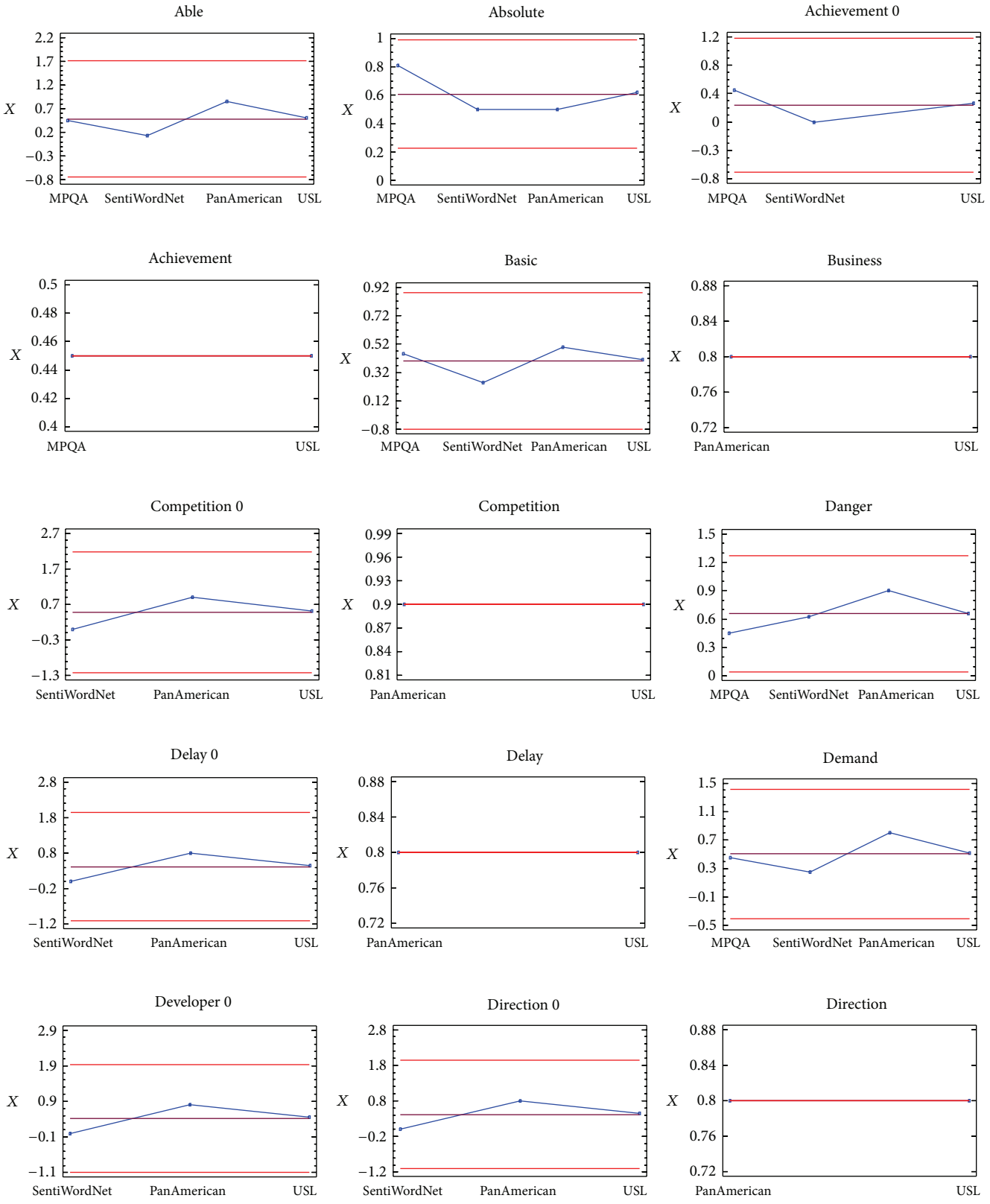

Figure 5: Continued.

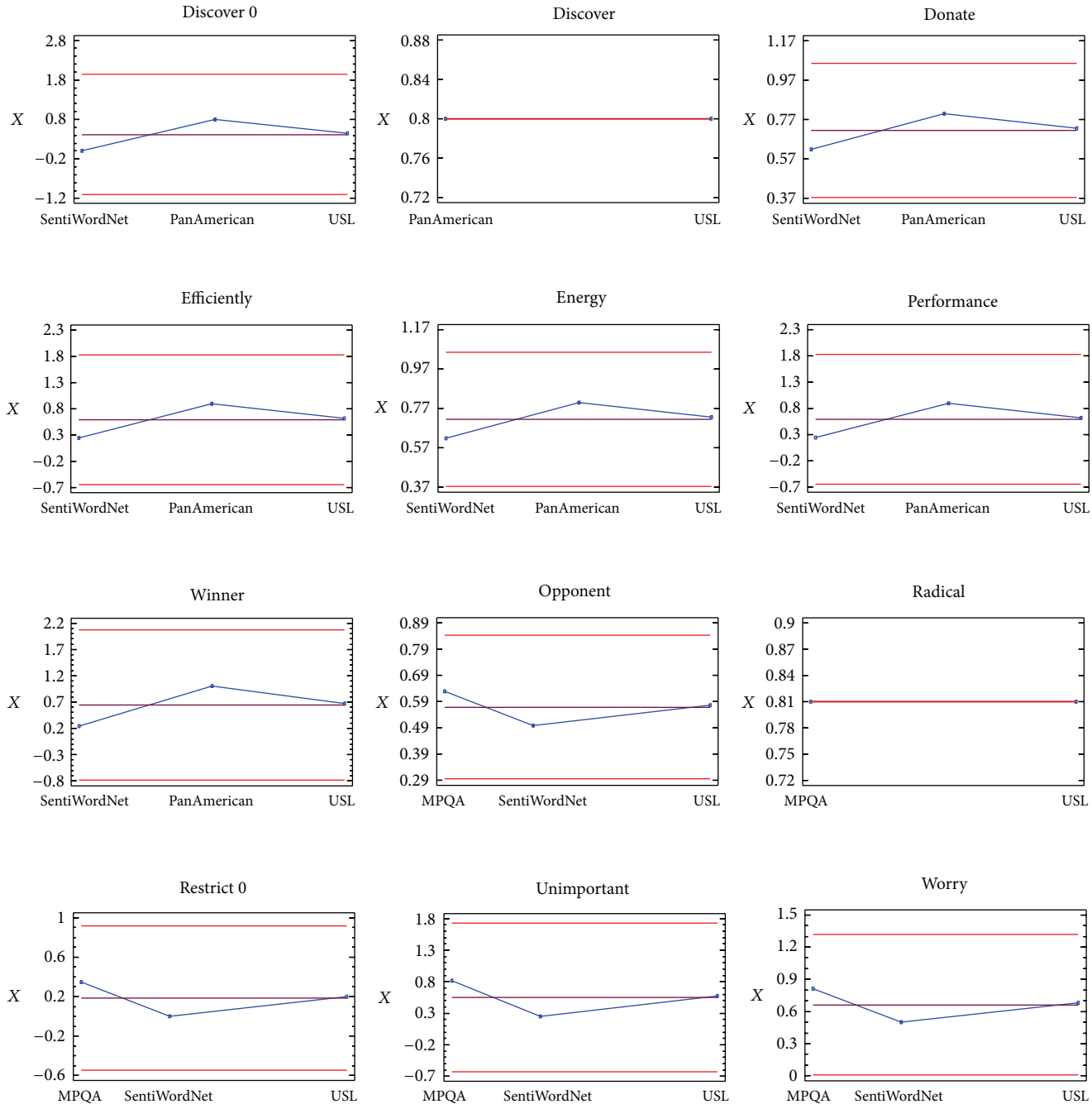

Figure 5: Strength of polarity for each lexical entry in Table 3 for the Unified Sentiment Lexicon, the SentyWordNet, PanAmerican, and MPQA sentiment lexicons.

The tests were performed on a CPU processor Intel XEON Hexa Core 2.50 Ghz, with 64 GiB of Ram and four GPUs: one Quadro 600 card with 96 CUDA Cores and three Tesla C2075 cards with 448 CUDA cores.

The experiment was conducted with 10, 50, 100, 500, 1,000, 5,000, and 95,430 lexical entries, respectively. As a result, the time was reducted to 3 times for the first set of data as show in Figure 6.

6. Discussion

It is clear that there must be progress in the task of assessing the quality of the linguistic resources. The above mentioned task takes working time (hours) on the part of expert linguists, but this work is needed if quality improvement is desired. Establishing priority criteria and stages for evaluating existing resources could help to implement this work

Table 3: A partial view of the USL strength of polarity for the English cluster.

Id Lexical	Intersection	BingLiu	MPQA	SentiWordNet	PanAmerican	α	β	γ	δ	ϵ	ζ	USL
1	Able	X	+0.45	+0.125	+0.85	X	+0.4122	+0.0825	+0.7505	+1.2452	+2.459	**+0.5064**
2	Absolute	X	+0.81	+0.5	+0.5	X	+0.7419	+0.33	+0.4415	+1.5134	+2.459	**+0.6154**
3	Achievement	X	+0.45	0	X	X	+0.4122	0	X	+0.4122	+1.576	**+0.2615**
4	Basic	X	+0.45	+0.25	+0.5	X	+0.4122	+0.165	+0.4415	+1.0187	+2.459	**+0.4142**
5	Business	X	X	+0.8	+0.85	X	X	X	+0.7064	+0.7064	+0.883	**+0.8**
6	Competition	X	X	0	+0.9	X	X	0	+0.7947	+0.7947	+1.543	**+0.515**
7	Danger	X	−0.45	−0.625	−0.9	X	−0.4122	−0.4125	−0.7947	−1.6194	−2.459	**−0.6585**
8	Delay	X	X	0	−0.8	X	X	0	−0.7064	−0.7064	−1.543	**−0.4578**
9	Demand	X	−0.45	−0.25	−0.8	X	−0.4122	−0.165	−0.7064	−1.2836	−2.459	**−0.5220**
10	Developer	X	X	0	+0.8	X	X	0	+0.7064	+0.7064	+1.543	**+0.4578**
11	Direction	X	X	0	∗0.8	X	X	0	∗0.7064	∗0.7064	∗1.543	**∗0.4578**
12	Discover	X	X	0	+0.8	X	X	0	+0.7064	+0.7064	+1.543	**+0.457**
13	Donate	X	X	+0.625	+0.8	X	X	+0.4125	+0.7064	+1.1189	+1.543	**+0.725**
14	Efficiently	X	X	+0.25	+0.9	X	X	+0.165	+0.7947	+0.9597	+1.543	**+0.6219**
15	Energy	X	X	+0.625	+0.8	X	X	+0.4125	+0.7064	+1.1189	+1.543	**+0.7251**
16	Performance	X	X	+0.25	+0.9	X	X	+0.165	+0.7947	+0.9597	+1.543	**+0.6219**
17	Winner	X	X	+0.25	+1	X	X	+0.165	+0.883	+1.048	+1.543	**+0.679**
18	Opponent	X	−0.63	−0.5	X	X	−0.5770	−0.33	X	−0.9070	−1.576	**−0.575**
19	Radical	X	−0.81	0	X	X	−0.7419	0	X	−0.7419	−1.576	**−0.4707**
20	Restrict	X	−0.35	0	X	X	−0.3206	0	X	−0.3206	−1.576	**−0.2034**
21	Unimportant	X	−0.81	−0.25	X	X	−0.7419	−0.165	X	−0.906	−1.576	**−0.5754**
22	Worry	X	−0.81	−0.5	X	X	−0.7419	−0.33	X	−1.071	−1.576	**−0.6801**
n	word$_n$	v_n	x_n	y_n	z_n	α_n	β_n	γ_n	δ_n	ϵ_n	ζ_n	**usl**$_n = \epsilon_n/\zeta_n$

Table 4: A partial view of the USL strength of polarity for the Spanish cluster.

Id Lexical	Intersection	STSL	PanAmerican	α	β	ϵ	ζ	USL
1	Agresividad	X	+0.85	X	+0.7505	+0.7505	+0.883	**+0.85**
2	Amigo	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
3	Absurdo	X	−0.9	X	−0.9	−0.9	−0.883	**−1.01**
4	Acariciar	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
5	Cansado	X	−0.8	X	−0.7064	−0.7064	−0.883	**−0.8**
6	Caminata	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
7	Active	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
8	Celebrar	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
9	Emocionar	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
10	Descanso	X	+0.8	X	+0.7064	+0.7064	+0.883	**+0.8**
11	Equivocarse	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
12	Admirado	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
13	Aprovechar	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
14	Ayudar	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
15	Vencer	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
16	Sudar	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
17	Merecer	X	+1	X	+0.883	+0.883	+0.883	**+1**
18	Deleite	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
19	Eficaz	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
20	Empeorar	X	−0.9	X	−0.7947	−0.7947	−0.883	**−0.9**
21	Desilusion	X	−0.9	X	−0.7947	−0.7947	−0.883	**−0.9**
22	Desgracia	X	−0.9	X	−0.7947	−0.7947	−0.883	**−0.9**
23	Dudar	X	−0.9	X	−0.7947	−0.7947	−0.883	**−0.9**
24	Dolor	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
25	Erroneo	X	−0.9	X	−0.7947	−0.7947	−0.883	**−0.9**
26	Economico	X	+0.9	X	+0.7947	+0.7947	+0.883	**+0.9**
27	Atacar	X	−0.9	X	−0.7947	−0.7947	−0.883	**−0.9**
n	word$_n$	x_n	y_n	α_n	β_n	ϵ_n	ζ_n	**usl**$_n = \epsilon_n/\zeta_n$

```
<owl:NamedIndividual rdf:ID="variedad">
    <rdf:type rdf:resource="#PositiveNouns"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:ID="corrupcion">
    <rdf:type rdf:resource="#NegativeNouns"/>
  </owl:NamedIndividual>
<owl:NamedIndividual rdf:ID="generoso">
    <rdf:type rdf:resource="#PositiveAdjectives"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:ID="obeso">
    <rdf:type rdf:resource="#NegativeAdjectives"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:ID="tranquilizar">
    <rdf:type rdf:resource="#PositiveVerbs"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:ID="increpar">
    <rdf:type rdf:resource="#NegativeVerbs"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:ID="a_faltar">
    <rdf:type rdf:resource="#NegativeAdverbs"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:ID="por_fortuna">
    <rdf:type rdf:resource="#PositiveAdverbs"/>
  </owl:NamedIndividual>
```

Algorithm 3

Table 5: A partial view of the USL strength of polarity for the Chinese Cluster.

Id Lexical	Intersection	NTUSD	SentiWordNet	α	β	ϵ	ζ	USL
1	(courageous)	X	+0.375	X	+0.3311	+0.3311	+0.883	**0.375**
2	(content)	X	+0.45	+0.125	+0.85	X	+0.4122	**+0.5064**
3	(agreement)	X	0	X	+0.9	+0.9	+0.883	**+1.019**
4	(perfection)	X	−0.5	X	−0.4415	−0.4415	−0.883	**−0.5**
5	(philosopher)	X	0	X	0	0	0.883	**0.0**
6	(difficulty)	X	−0.5	X	−0.4415	−0.4415	−0.883	**−0.5**
7	(sublime)	X	+0.625	X	+0.5518	+0.5518	+0.883	**+0.625**
8	(fabulous)	X	+0.875	X	+0.772	+0.772	+0.883	**+0.875**
9	(endeavour)	X	0	X	0	0	0.883	**0**
10	(good)	X	+0.625	X	+0.5518	+0.5518	+0.883	**+0.625**
11	(welcome)	X	+0.5	X	+0.4415	+0.4415	+0.883	**+0.5**
12	(praise)	X	0	X	0	0	0.883	**0**
13	(severe)	X	−0.625	X	−0.5528	−0.5518	0.883	**−0.625**
n	word$_n$	x_n	y_n	α_n	β_n	ϵ_n	ζ_n	**usl**$_n = \epsilon_n/\zeta_n$

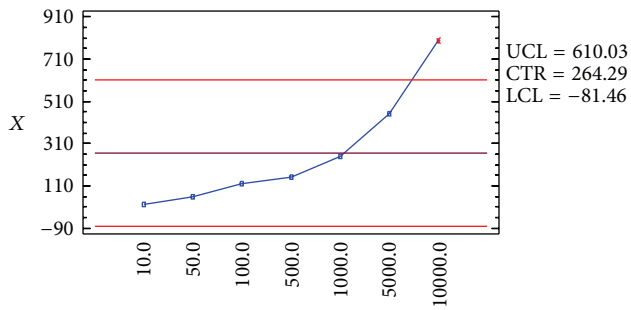

Figure 6: Number of gain at different scales of lexical entries.

realistically. If multilinguality is an aim then lexical resources in more languages need to be developed.

7. Conclusion and Future Work

We show that it is possible to unify the sentiment lexicons available on the web and align and expand them automatically. Our USL approach reuses the research work carried out by universities such as Illinois, Pittsburg, The Institute of Information Science, Taiwan, Istituto di Scienza e Tecnologie dellÍnformazione, and Universidad Politécnica de Madrid.

These sentiment lexicons have been essential for the implementation of the Unified Sentiment Lexicon. Our aim is to establish the USL as a standard that could be enriched and used by the whole community in the future.

The results of the USL approach are (a) the Unified Sentiment Lexicon and (b) OntoLexicon. Using parallel processing for the calculation of strength of polarity for each lexical entry, the USL approach accelerated the runtime by 300%. USL approach avoids hard disk operations and distributes the calculation of the USL metric over 1536 processors doing operations directly in GPU memory. The unification was carried out by means of providing a single strength of polarity for each lexical entry; this value must be present at least in one intersection in two or more sentiment lexicons in the same cluster. Compared with previous work, the major contributions of this paper are the following.

(i) A knowledge base of four sentiment lexicons (Bing Liu sentiment lexicon; MPQA; NTU Sentiment Dictionary; and SentiWordNet) has been unified automatically, grouped into three clusters—English, Spanish, and Chinese. In the final version Portuguese was not included because there are not enough sentiment lexicons available.

(ii) The USL approach computes a unified strength of polarity which was validated by experts.

(iii) The USL were expanded, with two additional sentiment lexicons that were developed by our research group, Communication in Specialized Domains: Pan-American sentiment lexicon and Spanish Travel Subjective Lexicon.

(iv) The task of strength of polarity unification uses parallel processing to compute each lexical entry with GPUs.

(v) USL computing time was accelerated 300% in data processing.

(vi) The robustness of the Unified Sentiment Lexicon was proven with 35201 positive, 38200 neutral, and 22029 negative lexical entries.

(vii) A uniform knowledge representation of the sentiment lexicon was made with OntoLexicon in OWL language.

Future work will involve proving the second research question: is it possible to transform a Unified Sentiment Lexicon into a generative lexicon based on a core ontology? We already have the core ontology; however, we need to transform this into a more generative lexicon. In addition, we need to extend the USL to other languages and domains, with the aim of having a unified linguistic resource to facilitate the task of subjective annotation both on the web and out of it.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

The authors are grateful to the Sciences Research Council (CONACYT) for funding this research project and they also thank NVIDIA for donating a TeslaK20 card.

References

[1] http://www.census.gov/hhes/socdemo/language/data/acs/appendix.html.

[2] M. Lewis, G. F. S. Paul, and C. D. Fennig, *Ethnologue: Languages of the World*, SIL International, Dallas, Texas, USA, 2013.

[3] M. R. Villarreal, "Corpus de blogs de viajes: análisis lingüístico para el reconocimiento de la valoración de la información," in *Proyecto Fin De Carrera. E. U. I. T. De Telecomunicación*, Universidad Politécnica de Madrid, 2009, (Spanish).

[4] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald, "The viability of web-derived polarity lexicons," in *Proceedings of the Human Language Technologies Conference ofthe North American Chapter of the Association for Computational Linguistics (NAACL HLT '10)*, pp. 777–785, Stroudsburg, Pa, USA, June 2010.

[5] A. Trujillo, "Automating the lexicon: research and practice in a multilingual environment," *Natural Language Engineering*, vol. 2, no. 3, pp. 277–285, 1996.

[6] E. Vegas, "Breadth and depth of semantic lexicons," *Computational Linguistics*, vol. 26, no. 4, pp. 652–656, 2000.

[7] S. Staab and R. Studer, *Handbook on Ontologies*, Springer, Berlin, Germany, 2nd edition, 2009.

[8] W. Abramowicz, Ed., *Business Information Systems*, vol. 87 of *Lecture Notes in Business Information Processing*, Springer, Berlin, Germany, 2011.

[9] A. Pak, *Automatic, adaptive, and applicative sentiment analysis [Ph.D. thesis]*, Université Paris-Sud, Orsay, France, 2012.

[10] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-basedmethods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[11] B. Heerschop, A. Hogenboom, and F. Frasincar, "Sentiment lexicon creation from lexical resources," *BIS*, pp. 185–196, 2011.

[12] J. Pustejovsky, C. Havasi, J. Littman, A. Rumshisky, and M. Verhagen, "Towards a generative lexical resource: the brandeis semantic ontology," in *Proceedings of the 5th Language Resource and Evaluation Conference*, 2006.

[13] J. Pustejovsky, "The generative lexicon," *Computational Linguistics*, vol. 17, no. 4, pp. 409–441, 1991.

[14] S. Bergler, "Metonymy and metaphor—boundary cases and the role of a generative lexicon," in *Proceedings of the First International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland, 2001.

[15] R. Thomas Gruber, "toward principles for the design of ontologies used for knowledge sharing," in *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Substantial Revision of Paper Presented at The International Workshop on Formal Ontology, Kluwer Academic Publishers, 1993.

[16] A. Sheppard, *Programming GPUs. Oreilly and Associate Series*, O'Reilly Media, 2012.

[17] J. Yan, D. B. Bracewell, F. Ren, and S. Kuroiwa, "The creation of a chinese emotion ontology based on hownet," *Engineering Letters*, vol. 16, no. 1, pp. 166–171, 2008.

[18] J. Wiskin, D. Borup, S. Johnson et al., "Inverse scattering and refraction corrected reflection for breast cancer imaging," in

Medical Imaging: Ultrasonic Imaging, Tomography, and Therapy, vol. 7629 of *Proceedings of SPIE*, February 2010.

[19] L.-W. Ku and H.-H. Chen, "Mining opinions from the web: beyond relevance retrieval," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1838–1850, 2007.

[20] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the LREC*, 2010.

[21] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[22] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 168–177, ACM, New York, NY, USA, August 2004.

[23] B. Liu and Hu, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web (WWW '05)*, pp. 342–351, ACM, New York, NY, USA, 2005.

[24] J. R. Martin and P. R. White, *The Language of Evaluation. Appraisal in English*, Palgrave, 2005.

[25] D. Kirk and W. Hwu, "Programming massively parallel processors: a hands-on approach," Applications of GPU Computing Series, Elsevier Science, 2010.

[26] S. Cook, "CUDA programming: a developer's guide to parallel computing with GPUs," Applications of GPU Computing Series, Elsevier Science, 2012.

[27] S. P. Midkiff, "Automatic parallelization: an overview of fundamental compiler techniques," *Synthesis Lectures on Computer Architecture*, vol. 19, pp. 1–169, 2012.

[28] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with CUDA," *Queue*, vol. 6, no. 2, pp. 40–53, 2008.

[29] N. Wilt, *The Cuda Handbook: a Comprehensive Guide to Gpu Programming*, Prentice Hall, 2013.

[30] T. Brandvik and G. Pullan, "An accelerated 3D Navier-Stokes solver for flows in turbomachines," *Journal of Turbomachinery*, vol. 133, no. 2, Article ID 021025, 2010.

[31] B. G. Levine, J. E. Stone, and A. Kohlmeyer, "Fast analysis of molecular dynamics trajectories with graphics processing units-Radial distribution function histogramming," *Journal of Computational Physics*, vol. 230, no. 9, pp. 3556–3569, 2011.

[32] L. I. Barbosa and I. Alvarez-de-Mon-y-Rego, Towards a Unified Sentiment Lexicon (USL) based on Graphics Processing Units (GPUs), 2013.

[33] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "NVIDIA Tesla: a unified graphics and computing architecture," *IEEE Micro*, vol. 28, no. 2, pp. 39–55, 2008.

[34] J. Sanders and E. Kandrot, *CUDA by Example: an Introduction to General-Purpose GPU Programming*, Addison-Wesley Professional, 1st edition, 2010.

[35] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, (HLT/EMNLP '05)*, pp. 347–354, Stroudsburg, Pa, USA, October 2005.